"TO HIT, OR NOT TO HIT?"

IN SILICO MODELS OF *IN VITRO* NUCLEAR RECEPTOR TRANSACTIVATION

Michael-Rock Goldsmith[1], Daniel T. Chang[1], James R. Rabinowitz[1], Stephen B. Little[1], Raymond R. Tice[2]

[1]Office of Research & Development, US-Environmental Protection Agency, Research Triangle Park, NC USA, [2]National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC USA

www.epa.gov

Abstract

In silico models are used to inform and prioritize bioassay requirements. Here, we develop a model for predicting the ability of compounds to activate estrogen /androgen/ thyroid (EAT) receptors in HEK293 cell line transactivation assays, based on data obtained on 309 unique chemicals in ToxCast™ phase 1. In our model, a hit for either "E" OR "A" or "T" or any combination/permutation, in both or either agonist / antagonist mode according to the assays performed by the NIH Chemical Genomics Center (NCGC) is a "hit". A "non-hit" is a chemical that shows no activity on any of the same targets with the same ToxCast™ training set. A decision tree classifier was developed with our functional definition of *in vitro* transactivation "hit" described above as our 'class field', and a total of 205 assay endpoints (43 Novascreen *in vitro* assays, 150 molecular docking assays and 12 physicochemical or pharmacokinetic related quantitative structure activity relationships (QSAR)) were selected as 'tree fields'. Our classification tree was developed using two-fold cross validation using the binary decision tree classifier implemented in Molecular Operating Environment (Chemical Computing Group, Montreal Canada). The most valuable "bits" that provided information that could discriminate, hence enrich hits to non-hits were MDCK permeability, PDBID:1G54 (androgen receptor) and PDBID:2UW9 (pkb, or akt1) binding (*in silico* molecular docking). An interesting point is that the QSAR-predicted MDCK (Madin-Darby canine kidney cell) permeability was the most useful parameter in enriching a dataset. Another interesting feature is how the *in vitro* and additional QSPR and docking results were neglected, and only the information bits with maximum ability to enrich hit/no-hit dataset was preserved: a cell permeability property and two discrete docking targets; the resulting accuracy of the model has a misclassification rate of ~11% (based on the chemical space of the training chemicals). These methods assist in the interpretation of the actual assay and outline key determinants of transactivation *in vitro*: cell permeability, nuclear receptor binding, and specificity of competing targets; a nuclear translocatable "AR" protein over obligate cytoplasm confined non –specific protein. Applied to the NCGC Phase 1 Tox 21 2900 compound library, a total of 870 chemicals were "predictable" based on the domain of applicability, of which ~ 65 were considered to be "active" based on modeled receptor specificity and cell permeability. In summary, our *in silico* model provides us with a mechanism of identifying the most likely chemical candidates that have the proper trans-cellular absorption criteria in addition to identifying fundamental ligand/biomolecular interaction determinants encoded *in vitro* or *in silico* (docking) results that give rise to segregated and enriched data. **Disclaimer** *This document does not reflect US-EPA policy. The appearance or absence of product, services, companies, organizations, home pages or other websites in this research product does not imply any endorsement or non-endorsement thereof and reflects the views of the scientist alone.*

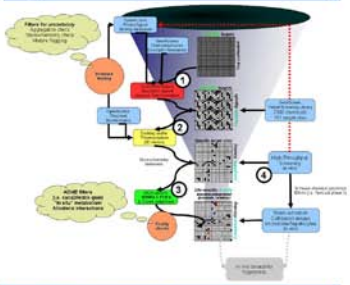
In Silico Models for Data Valuation

The computational toxicology paradigm addresses chemical risk management through a more data-driven, mechanistic and integrated "systems approach" and requires both *in silico* analysis as well as *in silico* derived information to close some data gaps. Although protein-binding, cell and tissue-based assays (*in vitro / in vivo* models) have been used to better inform *in vivo* toxicological outcomes, these models also require interpretation, understanding of domain of applicability, and limitations of the various methods; i.e. experimental boundary conditions.

In this sense the added value of *in silico* inquiry is both prioritization and interpretation. Interpretation of screening by knowledge mining allows one to define the expectations of the screening assays and the activity-based chemical/biological interactions they are attempting to probe; these approaches have added value to the information process by reducing surprise/erroneous interpretation of the assay.

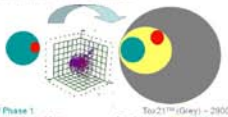
U.S. Environmental Protection Agency
Office of Research and Development

Integrated Workflow for Chemical Genomic Profiling



Data Sources and Domain of Applicability

The chemical space comparison (cartoon below) shows the training set (left) and the number of actives, compared to the *predictable space* of the test set (Tox21, right) and the number of predicted actives (based on decision tree classifiers). The PCA plot of PC1/PC2/PC3 are the normalized and de-correlated representation of Molecular Weight, LogPo/w and BrdR descriptors respectively, where purple spheres are Tox21 chemicals and red are ToxCast™ Phase I chemicals.

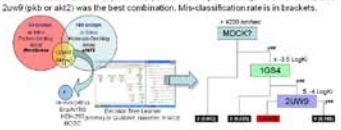


ToxCast™ Phase 1
EAT active in any mode (~ 10%)

Tox21™ (Grey) ~ 2900
Predictable = 870 + 320 (predicted ~65 actives)

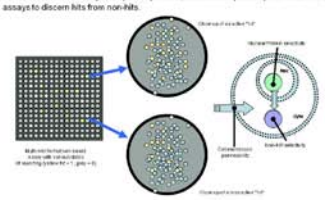
Knowledge Mining (KM) : Decision Tree Classifier

Of the various combinations used to get optimum separation between EAT modulator or non-modulator (i.e. hit (1) /no-hit (0)) the MDCK permeability, 1G54 (mutAR) and 2uw9 (pkb or akt2) was the best combination. Mis-classification rate is in brackets.



Interpretation of KM Analysis: Cell *Systems Models*

What is traditionally viewed only as a "hit/no-hit" scenario is phenomenologically disentangled as competing processes in a cellular *system*. The cartoon below depicts various transactivation reporter wells, but the very nature of magnitude versus viability is also drawn into question. At a molecular level of resolution we forget about the competing absorption, and competitive binding between nuclear translocatable protein (i.e. AR for instance) and other targets (the non-specific proteome). Our decision tree analysis has enabled us to discern key factors in HEK293 cell-line transactivation according to a specific protocol, and more importantly the value of key assays to discern hits from non-hits.



Future Work

To extend these methods of *in silico / in vitro* data fusion and subsequent knowledge mining to better prioritize and interpret *in vitro* cell-based assays we will develop a series of alternative models by dropping key selected receptors in each decision tree. Next, we will develop a consensus model and determine how this informs either the enrichment process or the phenomenological process, and extend these studies beyond Estrogen, Androgen and Thyroid Nuclear Receptors to other NRs evaluated in the Tox21 project. We will also establish a means to introduce "synthetic metabolism" into our model in order to effectively gauge impact on QSAR as a function of metabolic modifications of a parent chemical.

Acknowledgements

MRG and DTC thank Dr. Cecilia Tan (US-EPA, NERL, Exposure Dose Research Branch) for valuable discussion and review of this work, and Dr. David Dix (Deputy Director NCCT) for initiating this inter-laboratory ITR.